Stream: Internet Engineering Task Force (IETF)
RFC: 9816
Category: Informational
Published: July 2025
ISSN: 2070-1721
Authors: K. Patel, *Arrcus, Inc.*; A. Lindem, *Arrcus, Inc.*; S. Zandi; G. Dawra, *Google*; J. Dong, *Huawei Technologies*

RFC 9816
Usage and Applicability of BGP Link State (BGP-LS) Shortest Path First (SPF) Routing in Data Centers

Abstract

This document discusses the usage and applicability of BGP Link State (BGP-LS) Shortest Path First (SPF) extensions in data center networks utilizing Clos or Fat Tree topologies. The document is intended to provide simplified guidance for the deployment of BGP-LS SPF extensions.

Status of This Memo

This document is not an Internet Standards Track specification; it is published for informational purposes.

This document is a product of the Internet Engineering Task Force (IETF). It represents the consensus of the IETF community. It has received public review and has been approved for publication by the Internet Engineering Steering Group (IESG). Not all documents approved by the IESG are candidates for any level of Internet Standard; see Section 2 of RFC 7841.

Information about the current status of this document, any errata, and how to provide feedback on it may be obtained at https://www.rfc-editor.org/info/rfc9816.

Table of Contents

1. Introduction

This document complements [RFC9815] by discussing the applicability of the BGP Link State (BGP-LS) Shortest Path First (SPF) technology in a simple and fairly common deployment scenario, which is described in Section 3.

Section 4 describes the reasons for BGP modifications for such deployments.

Section 5 covers the BGP SPF protocol enhancements to BGP to meet these requirements and their applicability to data center [Clos] networks.

2. Recommended Reading

This document assumes knowledge of existing data center networks and data center network topologies [Clos]. This document also assumes knowledge of data center routing protocols such as BGP [RFC4271], BGP-LS SPF [RFC9815], and OSPF [RFC2328] [RFC5340] as well as data center Operations, Administration, and Maintenance (OAM) protocols like the Link Layer Discovery Protocol (LLDP) [RFC4957] and Bidirectional Forwarding Detection (BFD) [RFC5880].

3. Common Deployment Scenario

Within a data center, servers are commonly interconnected using the Clos topology [Clos]. The Clos topology is fully non-blocking, and the topology is realized using Equal-Cost Multipath (ECMP). In a multi-stage Clos topology, the minimum number of parallel paths in each tier is determined by the width of the stage as shown in Figure 1.

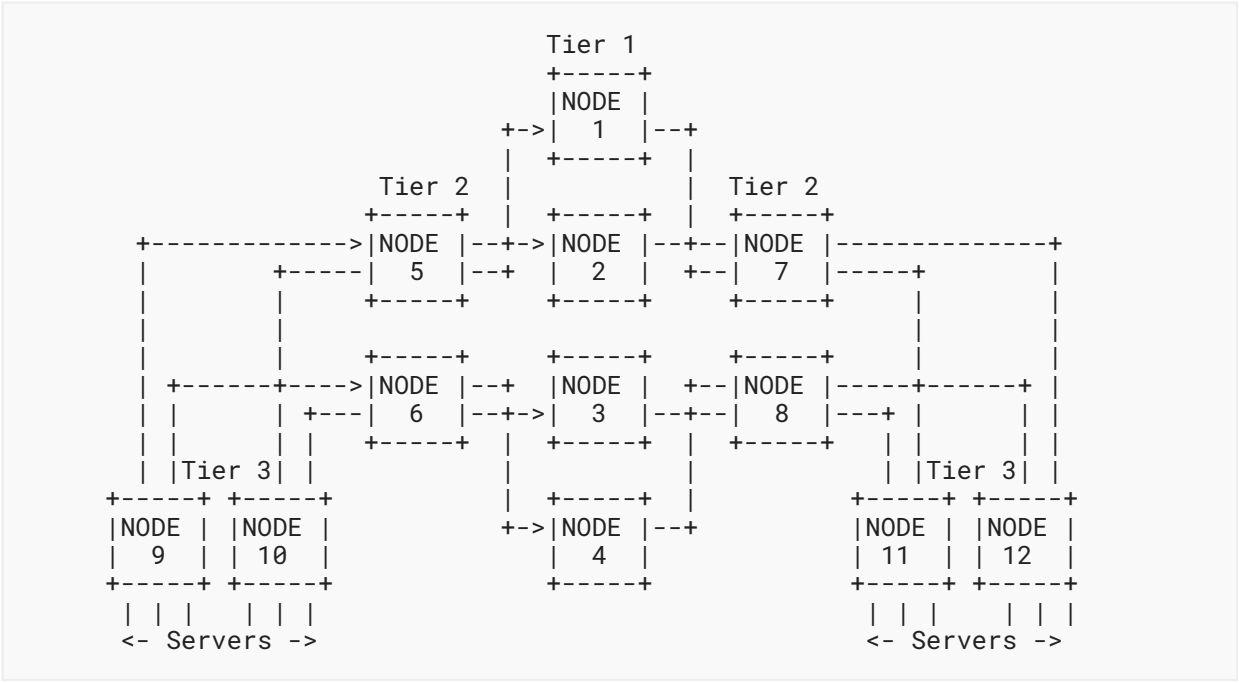

Figure 1: Illustration of the Basic Clos

- Tier 1 is comprised of Nodes 1, 2, 3, and 4
- Tier 2 is comprised of Nodes 5, 6, 7, and 8
- Tier 3 is comprised of Nodes 9, 10, 11, and 12

4. Justification for the BGP-SPF Extension

To simplify Layer 3 (L3) routing and operations, many data centers use BGP as a routing protocol to create both an underlay and an overlay network for their Clos topologies [RFC7938]. However, BGP is a path-vector routing protocol. Since it does not create a fabric topology, it uses hop-by-hop External BGP (EBGP) peering to facilitate hop-by-hop routing to create the underlay network and to resolve any overlay next hops. The hop-by-hop BGP peering paradigm imposes several restrictions within a Clos. It prohibits the deployment of route reflectors / route controllers as the EBGP sessions are congruent with the data path. The BGP best-path algorithm is prefix based, and it prevents announcements of prefixes to other BGP speakers until the best-path decision process has been performed for the prefix at each intermediate hop. These restrictions significantly delay the overall convergence of the underlay network within a Clos network.

The BGP SPF modifications allow BGP to overcome these limitations. Furthermore, using the BGP-LS Network Layer Reachability Information (NLRI) format allows the BGP SPF data to be advertised for nodes, links, and prefixes in the BGP routing domain [RFC9552] and used for SPF computations [RFC9815].

Additional motivation for deploying BGP-SPF is included in [RFC9815].

5. BGP-SPF Applicability to Clos Networks

With the BGP-SPF extensions [RFC9815], the BGP best-path computation and route computation are replaced with link-state algorithms such as those used by OSPF [RFC2328], both to determine whether a BGP-LS-SPF NLRI has changed and needs to be readvertised and to compute the BGP routes. These modifications will significantly improve convergence of the underlay while affording the operational benefits of a single routing protocol [RFC7938].

Data center controllers typically require visibility to the BGP topology to compute traffic-engineered paths. These controllers learn the topology and other relevant information via the BGP-LS address family [RFC9552], which is totally independent of the underlay address families (usually IPv4/IPv6 unicast). Furthermore, in usual BGP underlays, all the BGP routers will need to advertise their BGP-LS information independently. With the BGP-SPF extensions, controllers can learn the topology using the same BGP advertisements used to compute the underlay routes. Furthermore, these data center controllers can avail the convergence advantages of the BGP-SPF extensions. The placement of controllers can be outside of the forwarding path or within the forwarding path.

Alternatively, as each and every router in the BGP-SPF domain will have a complete view of the topology, the operator can also choose to configure BGP sessions in the hop-by-hop peering model described in [RFC7938] along with BFD [RFC5880]. In doing so, while the hop-by-hop peering model lacks the inherent benefits of the controller-based model, BGP updates need not be serialized by the BGP best-path algorithm in either of these models. This helps overall network convergence.

5.1. Usage of BGP-LS-SPF SAFI

Section 5.1 of [RFC9815] defines a new BGP-LS-SPF SAFI for announcement of the BGP-SPF link-state. The NLRI format and its associated attributes follow the format of BGP-LS for node, link, and prefix announcements. Whether the peering model within a Clos follows hop-by-hop peering as described in [RFC7938] or any controller-based or route-reflector peering, an operator can exchange BGP-LS-SPF SAFI routes over the BGP peering by simply configuring BGP-LS-SPF SAFI between the necessary BGP speakers.

The BGP-LS-SPF SAFI can also coexist with BGP IP Unicast SAFI [RFC4760], which could exchange overlapping IP routes. One use case for this is where BGP-LS-SPF routes are used for the underlay and BGP IP Unicast routes for VPNs are advertised in the overlay as described in [RFC4364]. The routes received by these SAFIs are evaluated, stored, and announced independently according to the rules of [RFC4760]. The tiebreaking of route installation is a matter of the local policies and preferences of the network operator.

Finally, as the BGP-SPF peering is done following the procedures described in [RFC4271], all the existing transport security mechanisms including those in [RFC5925] are available for the BGP-LS-SPF SAFI.

5.1.1. Relationship to Other BGP AFI/SAFI Tuples

Normally, the BGP-LS-SPF AFI/SAFI is used solely to compute the underlay and is given precedence over other AFI/SAFIs in route processing. Other BGP SAFIs, e.g., IPv6/IPv6 unicast VPN, would use the BGP-SPF computed routes for next-hop resolution.

5.2. Peering Models

As previously stated, BGP-SPF can be deployed using the existing peering model where there is a single-hop BGP session on each and every link in the data center fabric [RFC7938]. This provides for both the advertisement of routes and the determination of link and neighboring router availability. With BGP-SPF, the underlay will converge faster due to changes to the decision process that will allow NLRI changes to be advertised faster after detecting a change.

5.2.1. Sparse Peering Model

Alternately, BFD [RFC5880] can be used to swiftly determine the availability of links, and the BGP peering model can be significantly sparser than the data center fabric. BGP-SPF sessions only need to be established with enough peers to provide a biconnected graph. If Internal BGP (IBGP) is used, then the BGP routers at tier N-1 will act as route-reflectors for the routers at tier N.

The obvious usage of sparse peering is to avoid parallel BGP sessions on links between the same two routers in the data center fabric. However, this use case is not very useful since parallel L3 links between the same two BGP routers are rare in Clos or Fat Tree topologies. Additionally, when there are multiple links, they are often aggregated using Link Aggregation Groups (LAGs) at the link layer [IEEE.802.1AX] rather than at the IP layer. Two more interesting scenarios are described below.

In current data center topologies, there is often a very dense mesh of links between levels, e.g., leaf and spine, providing 32-way paths, 64-way paths, or more ECMPs. In these topologies, it is desirable not to have a BGP session on every link, and techniques such as the one described in Section 5.2.2 can be used to establish sessions on some subset of northbound links. For example, in a Spine/Leaf topology, each leaf router would only peer with a subset of the spines dependent on the flooding redundancy required to be reasonably certain that every node within the BGP-SPF routing domain has the complete topology.

Alternately, controller-based data center topologies are envisioned where BGP speakers within the data center only establish BGP sessions with two or more controllers. In these topologies, fabric nodes below the first tier, as shown in Figure 1 of [RFC7938], will establish BGP multi-hop sessions with the controllers. For the multi-hop sessions, determining the route to the controllers without depending on BGP would need to be through some other means, which is beyond the scope of this document. However, the BGP discovery mechanisms described in Section 5.5 would be one possibility.

5.2.2. Biconnected Graph Heuristic

With a biconnected graph heuristic, discovery of BGP SPF peers is assumed, e.g., as described in Section 5.5. In this context, "biconnected" refers to the fact that there must be an advertised Link NLRI for both BGP SPF peers associated with the link before the link can be used in the BGP SPF route calculation. Additionally, it is assumed that the direction of the peering can be ascertained. In the context of a data center fabric, the direction is either northbound (toward the spine), southbound (toward the Top-of-Rack (ToR) routers), or east-west (same level in the hierarchy). The determination of the direction is beyond the scope of this document. However, it would be reasonable to assume a technique where the ToR routers can be identified and the number of hops to the ToR is used to determine the direction.

In this heuristic, BGP speakers allow passive session establishment for southbound BGP sessions. For northbound sessions, BGP speakers will attempt to maintain two northbound BGP sessions with different routers. For east-west sessions, passive BGP session establishment is allowed. However, a BGP speaker will never actively establish an east-west BGP session unless it cannot establish two northbound BGP sessions.

BGP SPF sparse peering deployments not using this heuristic are possible but are not described herein and are considered out of scope.

5.3. BGP Spine/Leaf Topology Policy

One of the advantages of using BGP-SPF as the underlay protocol is that BGP policy can be applied at any level. For example, depending on the topology, it may be possible to aggregate or filter prefix advertisements using the existing BGP policy. In Spine/Leaf topologies, it is not necessary to advertise a BGP-LS Prefix NLRI received by leaf nodes from the spine back to other spine nodes. If a common Autonomous System (AS) is used for the spine nodes, this can easily be accomplished with EBGP and a simple policy to filter advertisements from the leaves to the spine if the first AS in the AS path is the spine AS.

In the figure below, the leaves would not advertise any NLRIs with AS 64512 as the first AS in the AS path.

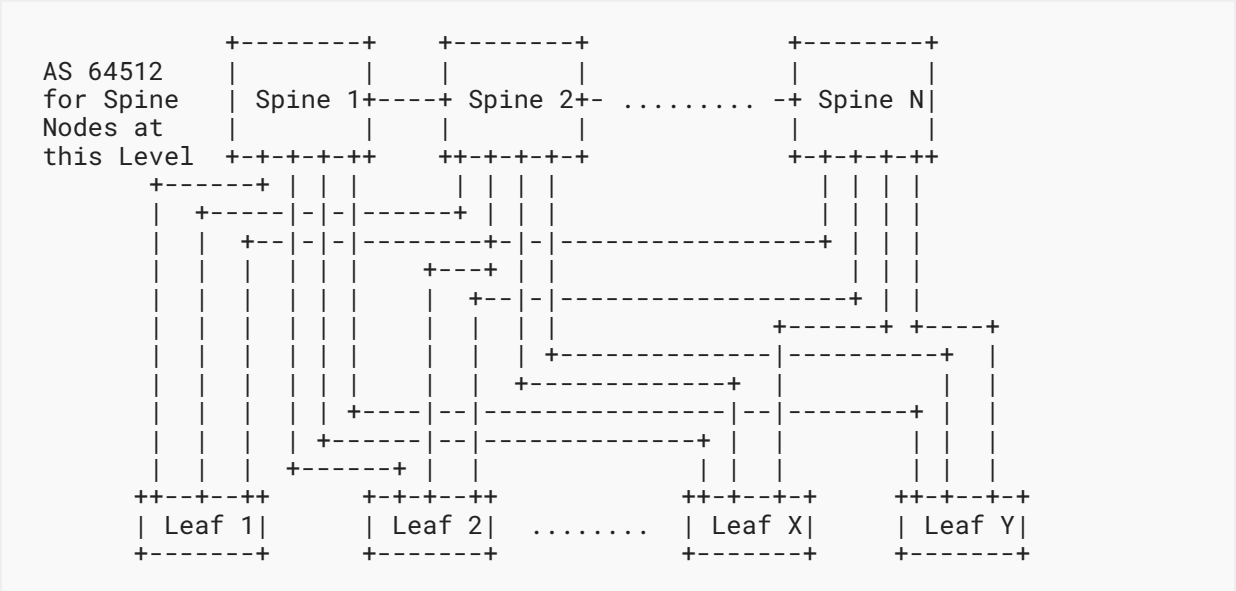

Figure 2: Spine/Leaf Topology Policy

5.4. BGP Peer Discovery Considerations

The basic functionality of peer discovery is to discover the address of a single-hop peer in the case where the peer address is not preconfigured. This is being accomplished today by using IPv6 Router Advertisements (RAs) [RFC4861] and assuming that a BGP session is desired with any discovered peer. Beyond the basic functionality, it may be useful to have the following information relating to the BGP session:

- The AS and BGP Identifier of a potential peer.
- Supported security capabilities, and for cryptographic authentication, the security capabilities and possibly a key chain [RFC8177] for use.
- A Session Policy Identifier, which is a group number or name used to associate common session parameters with the peer. For example, in a data center, BGP sessions with a ToR router could have different parameters than BGP sessions between leaf and spine nodes.

In a data center fabric, it is often useful to know whether a peer is southbound (towards the servers) or northbound (towards the spine or super-spine), e.g., see Section 5.2.2. One mechanism, without specifying all the details, might be for the ToR routers to be identified when installed and for the other routers in the fabric to determine their level based on the distance from the closest ToR router.

If there are multiple links between BGP speakers or the links between BGP speakers are unnumbered, it is also useful to be able to establish multi-hop sessions using the loopback addresses. This will often require the discovery protocol to install one or more routes toward the potential peer loopback addresses prior to BGP session establishment.

Finally, a simple BGP discovery protocol may be used to establish a multi-hop session with one or more controllers by advertising connectivity to one or more controllers.

5.5. BGP Peer Discovery

5.5.1. BGP IPv6 Simplified Peering

To conserve IPv4 address space and simplify operations, BGP-SPF routers in Clos / Fat Tree deployments can use IPv6 addresses as the peer address. For IPv4 address families, IPv6 peering as specified in [RFC8950] can be deployed to avoid configuring IPv4 addresses on router interfaces. When this is done, dynamic discovery mechanisms, as described in Section 5.5, can be used to learn the global or link-local IPv6 peer addresses, and IPv4 addresses need not be configured on these interfaces. If IPv6 link-local peering is used, then configuration of IPv6 global addresses is also not required [RFC7404]. The Link Local/Remote Identifiers of the peering interfaces must be used in the Link NLRI as described in Section 5.2.2 of [RFC9815].

5.5.2. BGP-LS-SPF Topology Visibility for Management

Irrespective of whether or not BGP-SPF is used for route calculation, the BGP-LS-SPF route advertisements can be used to periodically construct the Clos / Fat Tree topology. This is especially useful in deployments where an Interior Gateway Protocol (IGP) is not used and the base BGP-LS routes [RFC9552] are not available. The resultant topology visibility can then be used for troubleshooting and consistency checking. This would normally be done on a central controller or other management tool that could also be used for fabric data path verification. The precise algorithms and heuristics, as well as the complete set of management applications, is beyond the scope of this document.

5.5.3. Data Center Interconnect (DCI) Applicability

Since BGP-SPF is to be used for the routing underlay and Data Center Interconnect (DCI) gateway boxes typically have direct or very simple connectivity, BGP external sessions would typically not include the BGP-LS-SPF SAFI.

6. Non-Clos / Fat Tree Topology Applicability

The BGP-SPF extensions [RFC9815] can be used in other topologies and avail the inherent convergence improvements. Additionally, sparse peering techniques may be utilized as described in Section 5.2. However, determining whether to establish a BGP session is more complex, and the heuristic described in Section 5.2.2 cannot be used. In such topologies, other techniques such as those described in [RFC9667] may be employed. One potential deployment would be the underlay for a Service Provider (SP) backbone where usage of a single protocol, i.e., BGP, is desired.

7. Non-Transit Node Capability

In certain scenarios, a BGP node wishes to participate in the BGP-SPF topology but never be used for transit traffic. These include situations where a server wants to make application services available to clients homed at subnets throughout the BGP-SPF domain but does not ever want to

be used as a router (i.e., carry transit traffic). Another specific instance is where a controller is resident on a server and direct connectivity to the controller is required throughout the entire domain. This can readily be accomplished using the BGP-LS-SPF Node NLRI Attribute SPF Status TLV as described in [RFC9815].

8. BGP Policy Applicability

Existing BGP policy such as prefix filtering may be used in conjunction with the BGP-LS-SPF SAFI. When BGP policy is used with the BGP-LS-SPF SAFI, BGP speakers in the BGP-LS-SPF routing domain will not all have the same set of NLRIs and will compute a different BGP local routing table. Consequently, care must be taken to assure that routing is consistent and that routes to unreachable destinations or routing loops do not ensue. However, this is no different than if classical BGP routing using the IPv4 and IPv6 address families were used.

9. IANA Considerations

This document has no IANA actions.

10. Security Considerations

This document introduces no new security considerations above and beyond those already specified in [RFC4271] and [RFC9815].

11. References

11.1. Normative References

[RFC9815] Patel, K., Lindem, A., Zandi, S., and W. Henderickx, "BGP Link State (BGP-LS) Shortest Path First (SPF) Routing", RFC 9815, DOI 10.17487/RFC9815, July 2025, <https://www.rfc-editor.org/info/rfc9815>.

11.2. Informative References

[Clos] Clos, C., "A Study of Non-Blocking Switching Networks", The Bell System Technical Journal, vol. 32, no. 2, pp. 406-424, DOI 10.1002/j.1538-7305.1953.tb01433.x, March 1953, <https://doi.org/10.1002/j.1538-7305.1953.tb01433.x>.

[IEEE.802.1AX] IEEE, "IEEE Standard for Local and Metropolitan Area Networks--Link Aggregation", IEEE Std 802.1AX-2020, DOI 10.1109/IEEESTD.2020.9105034, May 2020, <https://doi.org/10.1109/IEEESTD.2020.9105034>.

[RFC2328] Moy, J., "OSPF Version 2", STD 54, RFC 2328, DOI 10.17487/RFC2328, April 1998, <https://www.rfc-editor.org/info/rfc2328>.

[RFC4271] Rekhter, Y., Ed., Li, T., Ed., and S. Hares, Ed., "A Border Gateway Protocol 4 (BGP-4)", RFC 4271, DOI 10.17487/RFC4271, January 2006, <https://www.rfc-editor.org/info/rfc4271>.

[RFC4364] Rosen, E. and Y. Rekhter, "BGP/MPLS IP Virtual Private Networks (VPNs)", RFC 4364, DOI 10.17487/RFC4364, February 2006, <https://www.rfc-editor.org/info/rfc4364>.

[RFC4760] Bates, T., Chandra, R., Katz, D., and Y. Rekhter, "Multiprotocol Extensions for BGP-4", RFC 4760, DOI 10.17487/RFC4760, January 2007, <https://www.rfc-editor.org/info/rfc4760>.

[RFC4861] Narten, T., Nordmark, E., Simpson, W., and H. Soliman, "Neighbor Discovery for IP version 6 (IPv6)", RFC 4861, DOI 10.17487/RFC4861, September 2007, <https://www.rfc-editor.org/info/rfc4861>.

[RFC4957] Krishnan, S., Ed., Montavont, N., Njedjou, E., Veerepalli, S., and A. Yegin, Ed., "Link-Layer Event Notifications for Detecting Network Attachments", RFC 4957, DOI 10.17487/RFC4957, August 2007, <https://www.rfc-editor.org/info/rfc4957>.

[RFC5340] Coltun, R., Ferguson, D., Moy, J., and A. Lindem, "OSPF for IPv6", RFC 5340, DOI 10.17487/RFC5340, July 2008, <https://www.rfc-editor.org/info/rfc5340>.

[RFC5880] Katz, D. and D. Ward, "Bidirectional Forwarding Detection (BFD)", RFC 5880, DOI 10.17487/RFC5880, June 2010, <https://www.rfc-editor.org/info/rfc5880>.

[RFC5925] Touch, J., Mankin, A., and R. Bonica, "The TCP Authentication Option", RFC 5925, DOI 10.17487/RFC5925, June 2010, <https://www.rfc-editor.org/info/rfc5925>.

[RFC7404] Behringer, M. and E. Vyncke, "Using Only Link-Local Addressing inside an IPv6 Network", RFC 7404, DOI 10.17487/RFC7404, November 2014, <https://www.rfc-editor.org/info/rfc7404>.

[RFC7938] Lapukhov, P., Premji, A., and J. Mitchell, Ed., "Use of BGP for Routing in Large-Scale Data Centers", RFC 7938, DOI 10.17487/RFC7938, August 2016, <https://www.rfc-editor.org/info/rfc7938>.

[RFC8177] Lindem, A., Ed., Qu, Y., Yeung, D., Chen, I., and J. Zhang, "YANG Data Model for Key Chains", RFC 8177, DOI 10.17487/RFC8177, June 2017, <https://www.rfc-editor.org/info/rfc8177>.

[RFC8950] Litkowski, S., Agrawal, S., Ananthamurthy, K., and K. Patel, "Advertising IPv4 Network Layer Reachability Information (NLRI) with an IPv6 Next Hop", RFC 8950, DOI 10.17487/RFC8950, November 2020, <https://www.rfc-editor.org/info/rfc8950>.

[RFC9552] Talaulikar, K., Ed., "Distribution of Link-State and Traffic Engineering Information Using BGP", RFC 9552, DOI 10.17487/RFC9552, December 2023, <https://www.rfc-editor.org/info/rfc9552>.

[RFC9667] Li, T., Ed., Psenak, P., Ed., Chen, H., Jalil, L., and S. Dontula, "Dynamic Flooding on Dense Graphs", RFC 9667, DOI 10.17487/RFC9667, October 2024, <https://www.rfc-editor.org/info/rfc9667>.

Acknowledgements

The authors would like to thank Alvaro Retana, Yan Filyurin, Boris Hassanov, Stig Venaas, Ron Bonica, Mallory Knodel, Dhruv Dhody, Erik Kline, Éric Vyncke, and John Scudder for their reviews and comments.

Authors' Addresses

Keyur Patel
Arrcus, Inc.
2077 Gateway Pl
San Jose, CA 95110
United States of America
Email: keyur@arrcus.com

Acee Lindem
Arrcus, Inc.
301 Midenhall Way
Cary, NC 27513
United States of America
Email: acee.ietf@gmail.com

Shawn Zandi
Email: shafagh@shafagh.com

Gaurav Dawra
Google
Sunnyvale, CA
United States of America
Email: gdawra.ietf@gmail.com

Jie Dong
Huawei Technologies
No. 156 Beiqing Road
Beijing
China
Email: jie.dong@huawei.com